CATEGORICAL DATA ANALYSIS

ASSIGNMENT II

Solution Set

2.1 The statement conveys $P(-|C) = 1/4$, $P(+|\bar{C}) = 2/3$.

Sensitivity $= P(+|C) = 1 - P(-|C) = 3/4$

Specificity $= P(-|\bar{C}) = 1 - P(+|\bar{C}) = 1/3$

2.2

Disease	Test Result	
Status	+	−
Disease	0.80	0.20
No Disease	0.20	0.80

$$\theta = \frac{(0.80)(0.80)}{(0.20)(0.20)} = 16$$

2.3 The response variable is injury status (fatal, nonfatal).

Difference in Proportions

$= \widehat{\pi}_{111} - \widehat{\pi}_{112} = (1601/164{,}128) - (510/412{,}878)$

$= 0.008519$

Relative Risk

$= \widehat{\pi}_{111} / \widehat{\pi}_{112} = (1601/164{,}128) \div (510/412{,}878)$

$= 7.897$

Odds Ratio

$$= \frac{\hat{\pi}_{111} / \hat{\pi}_{211}}{\hat{\pi}_{112} / \hat{\pi}_{212}} = \frac{m_{11} m_{22}}{m_{12} m_{21}} = \frac{(1601)(412,368)}{(162,527)(510)}$$

$$= 7.965$$

For Florida vehicle accidents in 1988:

- the Difference in Proportions indicates a difference of 0.001183 between the probability of sustaining a fatal injury if not wearing a seat belt and the corresponding probability if wearing a seat belt.
- the Relative Risk indicates that the probability of sustaining a fatal injury if not wearing a seat belt is 7.897 times higher than the corresponding probability if wearing a seat belt.
- the Odds Ratio indicates that the odds of sustaining a fatal injury if not wearing a seat belt are 7.965 times higher than the corresponding odds if wearing a seat belt.

The relative risk and the odds ratio are approximately the same because $\hat{\pi}_{111}$ and $\hat{\pi}_{112}$ are both small: i.e., the prevalence of fatal injuries in both seat belt groups is low.

2.7

(a) Difference in Proportions

= 0.001304 − 0.000121 = 0.001183

Relative Risk

= 0.001304 / 0.000121 = 10.78

For women in the US over the age of 35:

- the Difference in Proportions indicates a difference of 0.001183 between the annual probability of dying from lung cancer if a current smoker and the corresponding probability if a nonsmoker.
- the Relative Risk indicates that the annual probability of dying from lung cancer if a current smoker is 10.78 times higher than the corresponding probability if a nonsmoker.

The relative risk is more meaningful than the difference in proportions because the response probabilities are both small.

(b) Odds Ratio

$$= \frac{0.001304 / (1-0.001304)}{0.000121 / (1-0.000121)} = 10.79$$

For women in the US over the age of 35, the Odds Ratio indicates that the annual odds of dying from lung cancer if a current smoker are 10.79 times higher than the corresponding odds if a nonsmoker.

The relative risk and the odds ratio take similar values because the response probabilities are both small.

2.12 AG Conditional Odds Ratios

$$\hat{\theta}_{AG(A)} = \frac{(512)(19)}{(313)(89)} = 0.349$$

$$\hat{\theta}_{AG(B)} = \frac{(353)(8)}{(207)(17)} = 0.803$$

$$\hat{\theta}_{AG(C)} = \frac{(120)(391)}{(205)(202)} = 1.133$$

$$\hat{\theta}_{AG(D)} = \frac{(138)(244)}{(279)(131)} = 0.921$$

$$\hat{\theta}_{AG(E)} = \frac{(53)(299)}{(138)(94)} = 1.222$$

$$\hat{\theta}_{AG(F)} = \frac{(22)(317)}{(351)(24)} = 0.828$$

AG Marginal Odds Ratio

$$\hat{\theta}_{AG} = \frac{(1198)(1278)}{(1493)(557)} = 1.842$$

The AG conditional odds ratios tend to indicate that females are favored in the admission process: for four of the departments (A, B, D, F), the odds of a female being admitted exceed the odds of a male being admitted.

The AG marginal odds ratio indicates that males are favored in the admission process: over all departments, the odds of a male being admitted exceed the odds of a female being admitted.

The paradox can be resolved by noting that females tend to apply more often than males to competitive departments where admission rates are low: namely, C, D, E, and F.

2.14

(a)

Gender (Z)	Race (X)	Murder Victim (Y) Yes	No
Male	Nonwhite	0.0263	0.9737
	White	0.0049	0.9951
Female	Nonwhite	0.0072	0.9928
	White	0.0023	0.9977

$$\hat{\theta}_{XY(1)} = \frac{(0.0263)(0.9951)}{(0.9737)(0.0049)} = 5.485$$

$$\hat{\theta}_{XY(2)} = \frac{(0.0072)(0.9977)}{(0.9928)(0.0023)} = 3.146$$

The conditional odds ratios indicate that for each gender, the odds of being a murder victim are higher for nonwhites than for whites.

The association is not homogeneous since $\hat{\theta}_{XY(1)} \neq \hat{\theta}_{XY(2)}$.

(b)

Race (X)	Murder Victim (Y) Yes	No
Nonwhite	0.01675	0.98325
White	0.00360	0.99640

$$\hat{\theta}_{XY} = \frac{(0.01675)(0.99640)}{(0.98325)(0.00360)} = 4.715$$

2.19

Since the row and column variables are both ordinal, Gamma provides an appropriate measure of association between the variables.

$$
\begin{aligned}
C = {} & 7\,(8+3+7+5+4+9+8+9+14) \\
& + 2\,(5+4+9+8+9+14) \\
& + 1\,(8+9+14) \\
& + 7\,(3+7+4+9+9+14) \\
& + 8\,(4+9+9+14) \\
& + 5\,(9+14) \\
& + 2\,(7+9+14) \\
& + 3\,(9+14) \\
& + 4\,(14) \\
= {} & 1508
\end{aligned}
$$

$$
\begin{aligned}
D = {} & 2\,(7+2+3) \\
& + 1\,(7+2+3+8+3+7) \\
& + 2\,(7+2+3+8+3+7+5+4+9) \\
& + 8\,(2+3) \\
& + 5\,(2+3+3+7) \\
& + 8\,(2+3+3+7+4+9) \\
& + 3\,(3) \\
& + 4\,(3+7) \\
& + 9\,(3+7+9) \\
= {} & 709
\end{aligned}
$$

$$\hat{\gamma} = \frac{C-D}{C+D} = +0.360$$

There appears to be a positive association of moderate strength between a husband's and a wife's rating of sexual fun.

2.20 Let X correspond to Defendant's Race, Y correspond to Death Penalty, and Z correspond to Victim's Race.

To analyze the data and demonstrate that Simpson's paradox occurs, we will compute sample conditional and marginal odds ratios.

Conditional Odds Ratios (at each level of Z):

$$\hat{\theta}_{XY(1)} = \frac{(19)(52)}{(132)(11)} = 0.680$$

$$\hat{\theta}_{XY(2)} = \frac{(0)(97)}{(9)(6)} = 0$$

The conditional odds ratios indicate that for each victim's race, the odds of receiving the death penalty are higher if the defendant's race is black than if the defendant's race is white.

Marginal Odds Ratio (collapsed over Z):

$$\hat{\theta}_{XY} = \frac{(19)\,(149)}{(141)\,(17)} = 1.181$$

The marginal odds ratio indicates that for both victim's races combined, the odds of receiving the death penalty are higher if the defendant's race is white than if the defendant's race is black.

Simpson's paradox is illustrated by the fact that $\hat{\theta}_{XY(1)}, \hat{\theta}_{XY(2)}$ are both less than one and yet $\hat{\theta}_{XY}$ exceeds one.

2.21

(a) Let + (−) indicate a positive (negative) test result, and D ($\bar{D}$) indicate presence (absence) of the disease.

$$P(D|+) = \frac{P(+|D)\,P(D)}{P(+|D)P(D) + P(+|\bar{D})P(\bar{D})} \quad \text{(Bayes' Rule)}$$

$$= \frac{\pi_1 \rho}{\pi_1 \rho + \pi_2 (1-\rho)}$$

(b)

$$P(D|+) = \frac{(.95)(.005)}{(.95)(.005) + (1-.95)(1-.005)}$$

$$= 0.0872$$

$$P(+ \text{ and } D) = P(+|D)\,P(D)$$
$$= (.95)\,(.005) = 0.00475$$

$$P(+ \text{ and } \bar{D}) = P(+|\bar{D})\,P(\bar{D})$$
$$= (1-.95)(1-.005) = 0.04975$$

$$P(- \text{ and } D) = P(-|D)\,P(D)$$
$$= (1-.95)\,(.005) = 0.00025$$

$$P(- \text{ and } \bar{D}) = P(-|\bar{D})\,P(\bar{D})$$
$$= (.95)(1-.005) = 0.94525$$

Note that most of the tested subjects will be truly HIV- and have a negative test result.

For the tested subjects who have a positive test result, the probability this positive test will be accompanied by a true HIV- status is over 10 times as high as the probability the test will be accompanied by a true HIV+ status.

2.24 (a)

m_{11}	m_{12}
m_{21}	m_{22}

Table A

$\Rightarrow$

m_{11}	m_{21}
m_{12}	m_{22}

Table B

Note that Table B results from interchanging rows with columns in Table A.

	Table A	Table B
Odds Ratio	$\frac{m_{11} m_{22}}{m_{12} m_{21}}$	$\frac{m_{11} m_{22}}{m_{21} m_{12}}$
Difference in Proportions	$m_{11}/(m_{11}+m_{12}) - m_{21}/(m_{21}+m_{22})$	$m_{11}/(m_{11}+m_{21}) - m_{12}/(m_{12}+m_{22})$
Relative Risk	$\frac{m_{11}/(m_{11}+m_{12})}{m_{21}/(m_{21}+m_{22})}$	$\frac{m_{11}/(m_{11}+m_{21})}{m_{12}/(m_{12}+m_{22})}$

(b)

m_{11}	m_{12}
m_{21}	m_{22}

Table A

cm_{11}	m_{12}
cm_{21}	m_{22}

Table B

Note that Table B results from multiplying the counts in column 1 of Table A by c.

	Table A	Table B
Odds Ratio	$\frac{m_{11} m_{22}}{m_{12} m_{21}}$	$\frac{(\cancel{c} m_{11}) m_{22}}{m_{12} (\cancel{c} m_{21})}$

Difference in Proportions	$m_{11}/(m_{11}+m_{12})$ $- m_{21}/(m_{21}+m_{22})$	$(cm_{11})/((cm_{11})+m_{12})$ $-(cm_{21})/((cm_{21}+m_{22})$
Relative Risk	$\dfrac{m_{11}/(m_{11}+m_{12})}{m_{21}/(m_{21}+m_{22})}$	$\dfrac{(cm_{11})/((cm_{11})+m_{12})}{(cm_{21})/((cm_{21})+m_{22})}$

m_{11}	m_{12}
m_{21}	m_{22}

Table A

cm_{11}	cm_{12}
m_{21}	m_{22}

Table B

Note that Table B results from multiplying the counts in row 1 of Table A by c.

	Table A	Table B
Odds Ratio	$\dfrac{m_{11}\,m_{22}}{m_{12}\,m_{21}}$	$\dfrac{(\cancel{c}m_{11})\,m_{22}}{(\cancel{c}m_{12})\,m_{21}}$
Difference in Proportions	$m_{11}/(m_{11}+m_{12})$ $- m_{21}/(m_{21}+m_{22})$	$(\cancel{c}m_{11})/((\cancel{c}m_{11})+(\cancel{c}m_{12}))$ $- m_{21}/(m_{21}+m_{22})$
Relative Risk	$\dfrac{m_{11}/(m_{11}+m_{12})}{m_{21}/(m_{21}+m_{22})}$	$\dfrac{(\cancel{c}m_{11})/((\cancel{c}m_{11})+(\cancel{c}m_{12}))}{m_{21}/(m_{21}+m_{22})}$

Note: Here, all three measures are invariant to the operation.

2.28 $\theta_{XY(1)} = \theta_{XY(2)}$

$$\frac{\pi_{111}\,\pi_{221}}{\pi_{121}\,\pi_{211}} = \frac{\pi_{112}\,\pi_{222}}{\pi_{122}\,\pi_{212}} \qquad (a)$$

$\theta_{YZ(1)} = \theta_{YZ(2)}$

$$\frac{\pi_{111}\,\pi_{122}}{\pi_{112}\,\pi_{121}} = \frac{\pi_{211}\,\pi_{222}}{\pi_{212}\,\pi_{221}} \qquad (b)$$

Note that we can obtain (b) from (a) by multiplying both sides of (a) by

$$\frac{\pi_{211}\,\pi_{122}}{\pi_{221}\,\pi_{112}}$$

Thus, (a) and (b) are equivalent.

2.36 (a)

π_{11}	π_{12}
π_{21}	π_{22}

Consider a subject pair (A, B).

Let A_{ij} denote the event that subject A falls in cell (i,j),
let B_{ij} denote the event that subject B falls in cell (i,j).

$$\begin{aligned}\pi_c &= P((A_{11} \text{ and } B_{22}) \text{ or } (A_{22} \text{ and } B_{11}))\\ &= P(A_{11})P(B_{22}) + P(A_{22})P(B_{11})\\ &= \pi_{11}\pi_{22} + \pi_{22}\pi_{11} = 2\pi_{11}\pi_{22}\end{aligned}$$

$$\begin{aligned}\pi_d &= P((A_{12} \text{ and } B_{21}) \text{ or } (A_{21} \text{ and } B_{12})) \\ &= P(A_{12})P(B_{21}) + P(A_{21})\,P(B_{12}) \\ &= \pi_{12}\pi_{21} + \pi_{21}\pi_{12} = 2\pi_{12}\pi_{21}\end{aligned}$$

$$Q = \frac{\pi_c - \pi_d}{\pi_c + \pi_d} = \frac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{\pi_{11}\pi_{22} + \pi_{12}\pi_{21}}$$

(b) $-\pi_{12}\pi_{21} \leq +\pi_{12}\pi_{21}$

$$\pi_{11}\pi_{22} - \pi_{12}\pi_{21} \leq \pi_{11}\pi_{22} + \pi_{12}\pi_{21}$$

$$\frac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{\pi_{11}\pi_{22} + \pi_{12}\pi_{21}} \leq +1$$

$$Q \leq +1$$

$$-\pi_{11}\pi_{22} \leq +\pi_{11}\pi_{22}$$

$$-\pi_{11}\pi_{22} - \pi_{12}\pi_{21} \leq \pi_{11}\pi_{22} - \pi_{12}\pi_{21}$$

$$(-1)(\pi_{11}\pi_{22} + \pi_{12}\pi_{21}) \leq \pi_{11}\pi_{22} - \pi_{12}\pi_{21}$$

$$-1 \leq \frac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{\pi_{11}\pi_{22} + \pi_{12}\pi_{21}}$$

$$-1 \leq Q$$

(c) When $\pi_{11} = 0$ or $\pi_{22} = 0$, we have $\pi_c = 0$ and $Q = +1$. When $\pi_{12} = 0$ or $\pi_{21} = 0$, we have $\pi_d = 0$ and $Q = -1$.

(d)

$$Q = \frac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{\pi_{11}\pi_{22} + \pi_{12}\pi_{21}}$$

$$= \frac{\dfrac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} - \dfrac{\pi_{12}\pi_{21}}{\pi_{12}\pi_{21}}}{\dfrac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} + \dfrac{\pi_{12}\pi_{21}}{\pi_{12}\pi_{21}}}$$

$$= \frac{\theta - 1}{\theta + 1}$$